Finding the Observed Information when Using Monte Carlo E M for Mixed Models with Partially Observed / Grouped Data

Ranjini Natarajan
School of Operations Research
Cornell University
Ithaca, New York 14853

Charles E. McCulloch
Biometrics Unit and Statistics Center
Cornell University
Ithaca, New York 14853

BU-1263-M **September 1994**

Finding the Observed Information when Using Monte Carlo E M for Mixed Models with Partially Observed / Grouped Data

Ranjini Natarajan
School of Operations Research
Cornell University
Ithaca, NY 14853

Charles E. McCulloch
Biometrics Unit and Statistics Center
Cornell University
Ithaca, NY 14853

Abstract

In this work, we develop a method to estimate the observed information matrix when using Monte Carlo E M, for a class of mixed models for partially observed/grouped data. We propose a Monte Carlo sequel to Louis' method [3]. Our method includes a Gibbs step to generate variates from the appropriate densities. We illustrate the computations involved through two examples.

1 Introduction

A computational drawback of the E M algorithm is that often the E step involves hefty, sometimes insurmountable calculations (e.g., high dimensional integration). For some problems, it may be feasible to perform these calculations using direct numerical integration [4], although for more complicated models, this might not be a computationally tractable option. Tanner [6] outlined a Monte Carlo E M algorithm, where the idea is to replace the integrals involved in the E step with a Monte Carlo estimate. We develop a Monte Carlo sequel to Louis' [3] method to estimate the observed information matrix within the M C E M framework. Although this approach works quite generally, we have worked out the details for a class of mixed models for partially observed/grouped data. By partially observed data, we refer to censored or truncated data; by grouped data we refer to ordered categorical data. Our method includes a Gibbs step to generate variates from the appropriate densities. The computations involved are illustrated through two examples.

In Section 2, we outline Louis' method and describe a Monte Carlo implementation of his method. In Section 3, we formulate the class of mixed models of interest and describe the computations involved. In Section 4, we apply the methods developed in Section 3 to probit normal regression and censored regression.

2 Louis' Method

In the usual E M terminology, we define Y to be the *latent/complete* data with probability density or mass function denoted by $[Y \mid \theta]$, where θ is the unknown parameter vector and [.] denote densities. However, we do not observe Y; instead we observe a measurable function of Y, namely, $W \sim [W \mid \theta]$. The goal of E M is to find the maximum likelihood estimate of θ based on the observed data W. The E M method is only attractive in situations where finding the complete data maximum likelihood estimator and the observed information matrix is straightforward, but the problem based on the observed data requires an iterative solution.

Define the set $\mathcal{R} = \{y : w(y) = w\}$, i.e., $\mathcal{R}$ is the set of complete data Y that could have led to the observed data W. Louis [3] proved that the observed information matrix $I_W(\theta)$ satisfies the following identity:

$$I_W(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} \ln[Y \mid \theta] \mid Y \in \mathcal{R}\right] - \mathrm{Var}\left[\frac{\partial}{\partial\theta} \ln[Y \mid \theta] \mid Y \in \mathcal{R}\right] \quad (1)$$

The first term in $I_W(\theta)$ is simply the conditional expected information matrix of the complete data Y and is typically easy to compute. Louis proved that the second term is the expected information of the conditional distribution of Y given that Y lies in the set $\mathcal{R}$. In some applications, it may be computationally intractable to calculate the expectations in (1). Tanner [6] suggested a Monte Carlo approach to Louis' method by replacing the expectations with a Monte Carlo estimate, in the following way:

1) Generate $y_1, y_2, \ldots, y_m \overset{iid}{\sim} [Y \mid Y \in \mathcal{R}, \theta]$, for m suitably large.

2) Replace the first term in $I_W(\theta)$ by $-\frac{1}{m}\sum_{i=1}^{m} \frac{\partial^2}{\partial\theta^2} \ln[y_i \mid \theta]$ etc.

We now formulate the model of interest and illustrate the computations involved.

3 The Model

We consider the standard analysis of variance model for variance components estimation:

$$Y = X\beta + \sum_{k=1}^{r} Z_k u_k + \epsilon \quad (2)$$

$$u_k \sim N_{q_k}(0, \sigma_k^2 I) \quad (3)$$

$$\epsilon \sim N_n(0, \sigma_e^2 I) \quad (4)$$

where $Y \in \Re^{n \times 1}$ is the data vector which is partially observed or completely unobserved. $X \in \Re^{n \times p}$ is the design matrix associated with the unknown fixed effects vector $\beta \in \Re^{p \times 1}$ and $Z_k \in \Re^{n \times q_k}$ is the incidence matrix corresponding to the random effects vector u_k, $(k = 1, \ldots, r)$. We use the random effects structure as a convenient way to model the correlation among Y. The parameters of interest are $\theta = (\beta, \sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2, \sigma_e^2)$.

We say a component of Y, Y_i is unobserved, if the only data information available is that it lies in some interval (a_i, b_i) where $-\infty \leq a_i < b_i \leq \infty$, and at least one of a_i, b_i is finite. Such applications arise when "experimental conditions or measuring devices permit sample points to be trapped only within specified limits" [1] as in censored or truncated data.

To put this model in the E M framework, we define the vector Y to be the complete data, since given Y, finding the maximum likelihood estimates and their standard errors is a normal linear regression problem, which is easy. We define the set $\mathcal{R} = \{Y_i : Y_i = y_i, i \in U;\ Y_i : a_i < Y_i < b_i, i \in C\}$ where C is the set of indices corresponding to the unobserved components of Y and U that for the observed components of Y.

The complete-data log likelihood is given by:

$$\ln[Y \mid \theta] \propto -\frac{1}{2}\ln|V| - \frac{1}{2}(Y - X\beta)' V^{-1}(Y - X\beta)$$

where $V = \sum_{k=0}^{r} \sigma_k^2 Z_k Z_k'$, $\sigma_0^2 = \sigma_e^2$ and $Z_0 = I_n$.

The first term in (1) is a matrix whose components require the calculation of expectations of the following form:

- $E[(Y - X\beta)]$
- $E[(Y - X\beta)' V^{-1} Z_k Z_k' V^{-1} Z_l Z_l' V^{-1}(Y - X\beta)]$, $k, l = 0, \ldots, r$

where all expectations are conditional on $Y \in \mathcal{R}$. The second term involves expectations of the following form:

- $E[(Y - X\beta)(Y - X\beta)']$
- $E[(Y - X\beta)' V^{-1} Z_k Z_k' V^{-1}(Y - X\beta)]$, $k = 0, \ldots, r$
- $E[(Y - X\beta)(Y - X\beta)' V^{-1} Z_k Z_k' V^{-1}(Y - X\beta)]$, $k = 0, \ldots, r$
- $E[(Y - X\beta)' V^{-1} Z_k Z_k' V^{-1}(Y - X\beta)(Y - X\beta)' V^{-1} Z_l Z_l' V^{-1}(Y - X\beta)]$, $k, l, = 0, \ldots, r$

where all expectations are conditional on $Y \in \mathcal{R}$. So, in order to obtain a Monte Carlo estimate of $I_W(\theta)$, we need to generate $y_1, y_2, \ldots, y_m \sim [Y \mid Y \in \mathcal{R}, \theta]$ and then replace the expectations above by sums. It is interesting to note that we do not need to compute the first two expectations above separately, since $E[(Y - X\beta)' A (Y - X\beta)] = \text{trace}(A\, E[(Y - X\beta)(Y - X\beta)'])$, for any matrix A.

The density $[Y \mid Y \in \mathcal{R}, \theta]$ is not trivial to generate from, since it is the density of a multivariate normal constrained to lie within a certain set $\mathcal{R}$. We propose the use of the Gibbs sampler to generate variates from this distribution.

3.1 The Gibbs Sampler

We now outline the use of the Gibbs sampler. In order to generate a sample of Y's from the conditional distribution of $[Y \mid Y \in \mathcal{R}, \theta]$, we only need to generate the unobserved components from their full conditional distributions:

$$[Y_i, i \in C \mid Y_j, j \neq i]$$

which is a univariate truncated normal distribution, using standard results on normal theory. More formally, we have:
Step 0) Obtain starting values for Y_i, $i \in C$.

Step 1) For each $i \in C$, calculate

$$\sigma^2_{i \mid (i)} = \text{Var}[Y_i \mid Y_j = y_j, j \neq i]$$

and the covariance $\beta_{i\,|\,(i)} = \mathrm{cov}\,(Y_i,\, Y_{(i)})$, where $Y_{(i)} = (Y_1,\, Y_2,\, \ldots, Y_{i-1},\, Y_{i+1},\, \ldots, Y_n)'$.

Step 2) For each $i \in C$, calculate

$$\begin{aligned} \mu_{i\,|\,(i)} &= E[Y_i \,|\, Y_j,\, j \neq i] \\ &= x_i\,\beta + \beta'_{i\,|\,(i)}\,(Y_{(i)} - X_{(i)}\,\beta) \end{aligned}$$

where $X_{(i)} = X$ with row i deleted and x_i is the ith row of X.

Step 3) Simulate Y_i, $i \in C$ from a truncated normal distribution with mean $\mu_{i\,|\,(i)}$ and standard deviation $\sigma_{i\,|\,(i)}$, truncated between $(a_i,\, b_i)$.

Repeat Steps 2 and 3 a large number of times, NREP to get $Y^{(1)}, \ldots, Y^{(NREP)}$. Discard a suitable number NBURN of the $Y^{(j)}$ from the beginning of the sequence and then retain every NSKIPth one. Ofcourse, we only need to run the Gibbs sequence one time to generate a sample from $[Y \,|\, Y \in \mathcal{R},\, \theta]$. The advantages of this Gibbs sampling approach are two-fold. Firstly, we only ever need to generate variates from univariate truncated normal distributions, and fast acceptance-rejection algorithms exist to generate from truncated distributions [5]. Secondly, most of the computational effort is expended in repeating Steps 2 and 3 a large number of times. Thus, complicated random effects structures have little impact on the computational time, because they only affect Step 1. We verify our results on two data sets to illustrate the feasibility of the computations.

4 Examples

4.1 Probit Normal Regression

We consider a latent variable genesis of the probit normal model for binary data by postulating the existence of an underlying/latent variable Y. We assume that Y satisfies the linear mixed model in (2 - 4), with the error variance $\sigma_e^2 = 1$, without loss of generality [2]. We observe a binary variable $W_i = I(Y_i > 0)$; i.e., an indicator of whether Y crosses a threshold of 0. An example of a situation where such a threshold model might be appropriate is with regard to the *financial health* of a firm. The observed variable is an indicator of whether the firm is bankrupt (1/0), while the underlying variable represents the true health of the firm. It is unimportant whether we actually believe in the underlying variable, or merely use it as a device to estimate the parameters in the model. The advantage of this threshold model is that it automatically lends itself to a data augmentation approach such as the E M algorithm.

It is easy to see that $\mathcal{R}$ is simply the intersection of n half-lines; if $W_i = 1$, then we consider the half-line $[0,\, \infty)$ while if $W_i = 0$, we consider $(-\infty,\, 0]$. Thus, in Step 3) of the Gibbs sampler, we generate Y_i from a normal distribution, truncated *above* 0 if $W_i = 1$ and truncated *below* 0 if $W_i = 0$. We numerically verified our results on the Weil data set [7]. This data set has a treatment and control group and a single nested random effect. The response is survival status of rats and the random effect is litter. The observed data is binary indicating survival/death, and we assume it arises from a true underlying variable in the following way: $W_{ijk} = I(Y_{ijk} > 0)$ where

$$\begin{aligned} Y_{ijk} &= \beta_i + u_{ij} + \epsilon_{ijk} \\ u_{ij} &\sim N(0,\, \sigma_i^2\, I) \\ \epsilon_{ijk} &\sim N(0,\, 1) \end{aligned}$$

where i indexes treatment/control, j indexes litter and k indexes the rat within the litter. So, β_i is the group mean on the latent scale and the u_{ij} are the random litter effects. The following table shows the estimates of the standard errors of the maximum likelihood estimates obtained by numerical integration (Gaussian quadrature with 20 points) and our approach.

Group		**SE (M L E)**	
		Numerical	*M C Louis*
Treatment	$\hat{\beta}_1$	0.309	0.304 (0.002)
	$\hat{\sigma}_1$	0.291	0.297 (0.008)
Control	$\hat{\beta}_2$	0.169	0.167 (0.007)
	$\hat{\sigma}_2$	0.301	0.302 (0.028)

The Monte Carlo estimate is the average of 35 independent runs and each run is based on a Gibbs sample of size 1500. The numbers in parenthesis are the standard errors of the Monte Carlo estimate. We can see that our estimates agree substantially with those obtained by numerical integration.

4.2 Censored Regression

We consider the case where some of the Y are right censored. This can occur when the response is a waiting time and a typical member of the population of physical or biological units is observed till an event of interest (or censoring) occurs. Such data arise in medical applications (time till the first tumor), reliability (repairable systems and software reliability) or labor economics (period of successive layoffs).

The observed data is the pair $(\min(Y_i,\ a_i),\ I\ (Y_i \leq a_i),\ i = 1,\ \ldots,n)$. The response vector Y is assumed to satisfy the mixed model in (2 - 4). To put this model in the E M framework, we define Y to be the complete data. It is easy to see that $\mathcal{R} = \{Y_i = y_i, i \in U, Y_i > a_i, i \in C\}$ where U is the set of indices of uncensored observations and C that for censored observations. Again, in Step 3) of the Gibbs sampler, we simply generate the censored Y_i from a normal distribution, truncated above a_i. We applied our method to a matched pairs skin graft data set analyzed by Petitt [4]. This data concerns the survival of closely and poorly matched skin grafts on the same person. The model postulated for the logarithm of of the i^{th} survival time on the j^{th} subject, denoted by Y_{ij} is:

$$\begin{aligned} Y_{ij} &= \mu + \beta_j + \gamma\, g_{ij} + \epsilon_{ij} \\ \beta_j &\sim N(0,\ \sigma_\beta^2) \\ \epsilon_{ij} &\sim N(0,\ \sigma^2) \end{aligned}$$

where β_j is a single nested individual effect, μ is the overall mean, γ is a fixed regression parameter and g_{ij} is an indicator variable (-1 for a poor match and +1 for a good match). There were 2 censored observations in this data set. We compared our results on the standard errors of the fixed effects parameters, with those obtained by Petitt and they are displayed below.

Parameter	**S E (M L E)**	
	Petitt	*M C Louis*
μ	0.15	0.149 (5.027e-05)
γ	0.082	0.086 (6.297e-05)

The Monte Carlo estimate is the average of 50 independent runs and each run is based on a Gibbs sample of size 2000.

5 Conclusion

In this paper, we develop a method to estimate the standard errors of the maximum likelihood estimates for a class of mixed models for incomplete data. Our approach is a valuable contribution to the existing literature on likelihood inference, since we are now able to make inferential statements in situations where it may not even be possible to compute the likelihood function with any reasonable degree of precision. In addition to the examples discussed here, we have implemented our method for the Ordinal Probit model, Tobit regression and obtained satisfactory results.

References

[1] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) "Maximum Likelihood Estimation from Incomplete Data via the E M algorithm", *J. R. Statist. Soc. B*, Vol 39, 1-38.

[2] Harville D. A., Mee R. W. (1984) "A Mixed-Model Procedure for Analyzing Ordered Categorical Data", *Biometrics*, Vol 40, 393-408.

[3] Louis, T. A. (1982) "Finding the Observed Information Matrix when Using the E M Algorithm", *J. R. Statist. Soc. B*, Vol 44, 226-233.

[4] Petitt, A. N. (1986) "Censored Observations, Repeated Measures and Mixed Effects Models: An Approach using the E M Algorithm and Normal Errors", *Biometrika*, Vol 73, 635-643.

[5] Robert C. P. (1991) "Simulation of Truncated Normal Variables", Technical Report No. 161, LSTA, University of Paris 6.

[6] Tanner M. A. (1993) "Tools for Statistical Inference", Second Edition, Springer-Verlag, NY, 1993.

[7] Weil, C. S. (1970) "Selection of the Valid Number of Sampling Units and Consideration of their Combination in Toxicological Studies Involving Reproduction, Teratogenesis, or Carcinogenesis", *Food and Cosmetic Toxicology*, Vol 8, 177-182.